arXiv:2303.01577v1 [cs.HC] 2 Mar 2023

DeepLens：自然语言处理模型中的分布外数据检测

Da Song*阿尔伯塔大学 埃德蒙顿，加拿大 AB 省
dsong4@ualberta.ca

Zhijie Wang*阿尔伯塔大学 埃德蒙顿，加拿大 AB 省
zhijie.wang@ualberta.ca

Yuheng Huang 阿尔伯塔大学 埃德蒙顿，加拿大 AB 省
yuheng18@ualberta.ca

Lei Ma 加拿大阿尔伯塔大学 东京大学，日本
ma.lei@acm.org

Tianyi Zhang 普渡大学 美国印第安纳州西拉斐特
tianyi@purdue.edu



图 1：DeepLens 是一个用于支持自然语言处理模型中分布外（OOD）数据检测的交互式系统。开发者可以通过动态调整阈值并观察图标阵列和 OOD 分数分布的变化来检测分布外问题。DeepLens 还通过聚类相似文本并将关键词可视化，帮助开发者探索分布外类型。为了理解分布外数据，开发者可以查看高亮关键词，并将其与分布内（ID）数据进行比较。

摘要

机器学习（ML）已广泛应用于自然语言处理（NLP）应用中。机器学习中的一个基本假设是，训练数据和真实世界数据应遵循相似的分布。然而，由于真实世界数据中的分布偏移，已部署的机器学习模型可能会面临分布外（OOD）问题。尽管已有许多算法被提出用于从文本语料库中检测分布外数据，但目前仍缺乏为机器学习开发者提供交互式工具支持的方案。在本研究中，我们提出了 DeepLens，这是一种帮助用户检测和探索海量文本语料库中分布外问题的交互式系统。借助文本聚类方法，用户可以在 DeepLens 中高效地探索不同的分布外类型。用户还可以通过检查经神经元激活分析高亮显示的显著词，深入分析特定文本。在一项包含 24 名参与者的被试内用户研究中，使用 DeepLens 的参与者能够准确找到近两倍的分布外问题类型，与不具备交互或可视化支持的 DeepLens 变体相比，信心高出 22%。

ACM分类体系概念

• **以人为中心的计算** → **交互式系统与工具**；• **计算方法** → **自然语言处理**。

关键词

交互式可视化、分布外检测、机器学习、自然语言处理

ACM参考文献格式：
Da Song、Zhijie Wang、Yuheng Huang、Lei Ma 和 Tianyi Zhang。2023。DeepLens：自然语言处理模型中的分布外数据交互式检测。载于《*2023*年*CHI* 计算机系统中的人为因素会议论文集（*CHI '23*）》，*2023*年*4*月*23*日至*28*日，德国汉堡。ACM，纽约州纽约，美国，17页。
https://doi.org/10.1145/3544548.3580741

1 引言

机器学习（ML）技术和机器学习模型在许多应用中表现出色，例如自动驾驶 [49]，虚拟助手 [8]，和医学诊断 [32]。现代机器学习技术通常假设训练数据和测试数据遵循相似的分布。然而，这一假设在现实世界中很难成立。相反，数据分布偏移和分布外（OOD）样本往往会导致已部署机器学习模型的表现下降

*两位作者对本研究贡献相同。

允许免费制作本作品全部或部分内容的数字或纸质副本，供个人或课堂使用，但前提是不得为营利或商业利益而制作或分发副本，且副本首页须附有此声明和完整引文。对于本作品中由作者以外的其他方拥有版权的组成部分，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、重新发表、上传至服务器或分发至列表，则须事先获得具体许可和/或支付费用。请向 permissions@acm.org 申请许可。
CHI '23，*2023*年*4*月*23*日至*28*日，德国汉堡© 2023 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-1-4503-9421-5/23/04...$15.00
https://doi.org/10.1145/3544548.3580741

[17, 21, 37]。表现下降还会进一步引发人们对机器学习模型可靠性和可信度的担忧。尤其是在医疗保健等高风险、安全关键应用中，无法预测分布外样本可能会造成严重后果 [32]。2013年，Google Flu Trends（GFT）模型未能预测流感季，其预测的流感季峰值比实际值低了140% [30]。导致这一失败的一个重要因素是，GFT 没有考虑到用户的搜索行为自 2012 [30] 以来发生了变化。在这一案例中，与2012年相比，数据分布在2013年发生了偏移，最终导致模型表现不佳。

为减轻分布外数据的影响，研究人员提出了许多分布外检测技术 [4, 21, 33–35, 44, 58]。对于给定的数据实例，这些技术通常会先计算一个特定分数（分布外分数），然后将其与预先设定的阈值进行比较，以判断该数据实例是否属于分布外数据。然而，在实践中，仅识别出分布外数据并不足以满足机器学习开发者的需求。例如，在前面 GFT 的例子中，即使已经发现存在分布外问题，开发者仍需深入分析分布外数据，弄清楚这些数据为何被判定为分布外数据，以及它们有哪些特征。这一过程十分耗时。尽管近期已有一项技术 [9]被提出以应对这一挑战，但它仅适用于图像数据，而不适用于文本数据。与人类更容易一眼浏览的图像相比 [15], ，阅读和理解文本数据需要投入更多认知努力。如果缺少适当的工具支持，开发者要大规模调查海量分布外文本数据，可能会面临困难且耗时。

本文探索了交互式工具如何帮助用户快速检测大型文本语料库中的分布外样本并了解其背景。我们介绍了 DeepLens，这是一种新颖的交互式系统，可帮助用户检测、探索并理解分布外问题。DeepLens 基于最大软最大概率（MSP）构建，这是一种广受欢迎的、基于校准的文本数据分布外检测方法 [4, 21]。为了帮助用户探索文本语料库中不同类型的分布外数据，DeepLens 首先按主题对相似的分布外数据进行聚类，然后将每个簇中的高频词以词云形式呈现，帮助用户检查并理解各簇的主题。当用户深入查看单个分布外实例时，DeepLens 会通过神经元激活分析方法突出显示每个实例中的显著词 [1]。通过这种方式，DeepLens 帮助用户无需仔细通读长文本，即可快速理解其内容。DeepLens 还将分布内数据与分布外数据并列呈现，帮助用户对二者进行比较和对照。

为了评估 DeepLens 的可用性和效率，我们对 24 名在机器学习和自然语言处理方面具有不同专业水平的程序员开展了被试内用户研究。我们通过在 DeepLens 中禁用聚类视图和高亮视图，创建了一个对比基线。结果表明，在四种不同的自然语言处理任务中，使用 DeepLens 的参与者能够发现更多类型的分布外数据。每位参与者使用 DeepLens 和基线工具发现的分布外类型数量的均值分别为 3.54 和 1.25（Welch's t-检验：$p < 0.0001$）。使用 DeepLens 的参与者对自己发现的机器学习模型中的分布外问题也更有信心。在7点李克特量表上，中位数分别为 6 和 5（Welch's t-检验：$p = 0.002$）。这些结果表明，在处理自然语言处理模型中的分布外问题时，DeepLens 可以显著提高机器学习开发者的工作效率。

总之，本文的主要贡献是 DeepLens，这是一个交互式系统，可帮助用户检测、探索并理解大型文本语料库中的分布外数据。我们已在 GitHub 上开源该系统 [1]。一项被试内用户研究证明，DeepLens 能够有效检测和分析各种自然语言处理任务中的不同类型分布外问题。

2 背景与相关工作

2.1 机器学习系统中的分布外问题

机器学习理论的一个基本假设是，训练数据和测试数据服从相似的分布 [42]。然而，模型部署后，遇到与训练数据分布外的真实世界数据并不少见。先前的研究表明，当输入分布外样本时，机器学习模型可能会以很高的置信度给出错误的预测结果 [14, 45]。当这些预测结果用于指导医学诊断等真实世界决策时，此类错误可能造成严重后果，例如将健康样本错误地归类为致病样本，反之亦然 [3, 50, 52]。

多年来，人们一直致力于理解机器学习系统中的分布外问题。Moreno-Torres 等人 [43] 提出了一个统一框架，用于分析分布偏移。给定一个分类任务 $X \rightarrow Y$，X 和 Y 的联合概率可以表示为 $p(y, x) = p(y|x)p(x)$。Moreno-Torres 等人 [43] 随后将分布外问题分为两类：(1) 协变量偏移，即输入分布 $p(x)$ 发生变化；(2) 概念偏移，即输入变量与类别变量之间的关系 $p(y|x)$ 发生变化。Arora 等人 [4] 进一步将这一分类体系扩展到自然语言处理任务。他们假设文本数据可以表示为背景特征（例如体裁），这些特征在不同标签间保持不变；以及语义特征（例如情感词），这些特征对预测任务具有区分性。因此，他们将背景特征的变化定义为背景偏移，将语义特征的分布变化定义为语义偏移。由于我们的工作聚焦于文本数据中的分布外问题，因此采用 Arora 等人 [4], 提出的分布外分类体系和术语。

2.2 分布外检测

机器学习领域关于分布外检测的研究成果十分丰富。大多数先前的工作都会为每个输入计算一个分布外分数，并使用阈值将分布内数据与分布外数据区分开来。Hendrycks 等人 [21] 首先提出了一种检测分布外样本的简单方法，是这一方向最早的尝试之一。他们将预测概率（即模型置信度）用作分布外问题的指标，预测概率越低，分布外分数就越高。然而，由于深度学习模型常常会"自信地"犯错， [45], 仅利用模型置信度会限制分布外检测的进一步改进。为了解决这一问题，近期有些工作提出训练校准模型，使其对分布外数据给出置信度较低的预测。校准模型可以通过数据增强 [22, 57, 62], 、对抗训练 [7, 10, 20], 和不确定性建模 [5, 40] 获得。另一类工作则通过利用其他指标进行分布外检测，来突破模型置信度的限制 [4, 33–35, 44, 58]。其中一项代表性工作是 ODIN [33], ，它使用温度缩放和

[1] https://github.com/momentum-lab-workspace/DeepLens

输入扰动来计算分布外分数。此外，还可以通过估计分布内分布，并衡量输入实例与分布内分布之间的距离来实现分布外检测 [26, 29, 31]。在这项工作中，我们基于 Arora 等人 [4]，提出的分布外算法开发了 DeepLens，该算法利用最大 softmax 概率（MSP）对文本语料库进行分布外检测。

迄今为止，大多数努力程度都投入到了提高分布外检测算法的准确率。然而，仅提供一个分布外分数和一份分布外样本列表，不足以让人类理解和推理分布外数据。DeepLens 通过提供一个交互式系统填补了这一空白，该系统帮助开发人员探索从大型文本语料库中检测到的分布外数据，并了解其特征。

2.3 用于分布外检测的交互式支持

在过去两年中，人们对为检测分布变化提供交互式工具支持的兴趣日益增长 [9, 46, 47, 59 61] [9], – 。

OoDAnalyzer 是一个用于分析图像数据中分布外问题的交互式系统。它提供了一种基于网格的可视化，以网格视图展示各个分布外图像。此外，它允许用户放大查看单个分布外实例，并突出显示对预测结果有显著贡献的图像部分。OoDAnalyzer 与 DeepLens 的主要区别在于，OoDAnalyzer 专注于图像数据，而 DeepLens 专注于文本数据。与文本数据相比，图像更容易一目了然地理解。因此，OoDAnalyzer 的界面设计不适用于文本语料库中的分布外分析。为填补这一空白，DeepLens 利用一种文本聚类方法，并突出显示单个文本文件中的显著词，以帮助用户探索和理解分布外实例。

数据漂移检测与分布外检测密切相关。Yeshchenko 等人提出了可视化漂移检测（VDD） [61], ，这是一套用于检测和分析业务流程漂移的可视化与交互系统。VDD 利用一组交互式图表，以随时间变化的方式呈现业务流程漂移（事件序列数据）。Wang 等人提出了 ConceptExplorer [59], ，这是一套用于分析多源时间序列数据中概念漂移的可视分析系统。Yang 等人提出了 DriftVis [60], ，这是一套用于分析流数据中概念漂移的可视分析系统。它利用增量高斯混合模型检测发生概念漂移的样本，并通过预测级可视化呈现目标模型的性能变化。DriftVis 专为概念漂移（即自然语言处理中的语义偏移）而设计，而 DeepLens 并未针对任何特定类型的分布偏移进行专门设计，因此可应用于两种偏移类型。

3 用户需求与设计原理

在本节中，我们首先基于文献综述分析机器学习从业者对交互式分布外检测的需求。然后，我们通过系统概述讨论所提出的系统如何支持这些需求。

3.1 机器学习中检测和诊断分布外问题的用户需求

为了了解机器学习从业者的需求，我们回顾了以往开展过分布外检测形成性研究 [9, 47, 60, 61], 、进行过用户研究 [46, 59], ，或讨论过处理分布外问题所面临挑战的研究 [38, 51]。基于这项回顾，我们总结出分布外检测的五项主要用户需求。***N1***：自动检测分布外数据。人工检查单个实例以识别数据分布偏移既耗时又繁琐 [59, 61]。Yeshchenko 等人 [61] 与行业合作伙伴密切合作后发现，工业从业者要求及时、准确地识别分布偏移。因此，DeepLens 应根据用户定义的标准自动检测分布外数据。Wang 等人 [59] 的专家评审也证实了在大型数据集中进行自动化分布外检测的必要性。

N2：理解样本为何被检测为分布外。近期研究 [38, 47, 59, 60] 表明，仅检测分布外样本是不够的。在实践中，机器学习从业者通常还希望了解这些样本为何处于分布外。例如，Yang 等人 [60] 访谈了四位机器学习从业者，发现相比于仅仅获得检测出的分布外样本，机器学习从业者更希望了解分布偏移发生的原因和位置。在另一项对四位数据科学家的访谈研究中，Palmeiro 等人 [47] 发现，数据科学家希望了解数据集的哪些部分包含数据偏移，以及数据偏移的模式。

N3：识别不同类型的分布外数据。不同的分布外实例可能具有不同的特征。因此，机器学习从业者希望对分布外数据进行分类，以更好地理解其共性和差异，从而制定更全面的策略来考虑数据偏移带来的影响 [9, 51, 59, 61]。例如，Wang 等人 [59] 强调，用户应能够区分不同类型的数据，并验证每种类型的分布偏移。在一项同时访谈机器学习开发者和机器学习用户的研究中，Chen 等人 [9] 报告称，机器学习开发者和机器学习用户都希望通过可视化方式探索不同类型的分布外样本及其关系，以减少需要检查的样本数量。

N4：比较分布外数据与分布内数据。Chen 等人 [9] 发现，将具有相同预测标签的分布外样本与分布内样本进行比较，有助于用户确认潜在的分布外问题。Olson 等人 [46] 对 60 名机器学习用户开展了一项用户研究，发现用户经常通过比较分布外样本与分布内样本来了解分布外样本的特征。

N5：从全局和局部视角探查分布外问题。在检查分布外问题时，用户往往会先探索不同类别的潜在分布外数据，再深入感兴趣的类别，将分布外样本与相似的分布内样本进行比较 [9]。Yeshchenko 等人 [61] 强调，应支持对分布外数据进行“下钻”和“上卷”分析，让用户能够以不同粒度灵活地调查分布外问题。

3.2 设计依据

为了支持**N1**，DeepLens采用基于校准的方法 [21]，自动检测大型文本语料库中的分布外数据。用户可以在分布视图（图 2 A◯）中查看测试数据中分布外实例所占的百分比，并通过阈值滑块调整将实例视为分布外的程度。为了帮助用户更好地理解为何某些实例被检测为分布外数据（**N2**），DeepLens支持用户并排比较一个分布外实例和

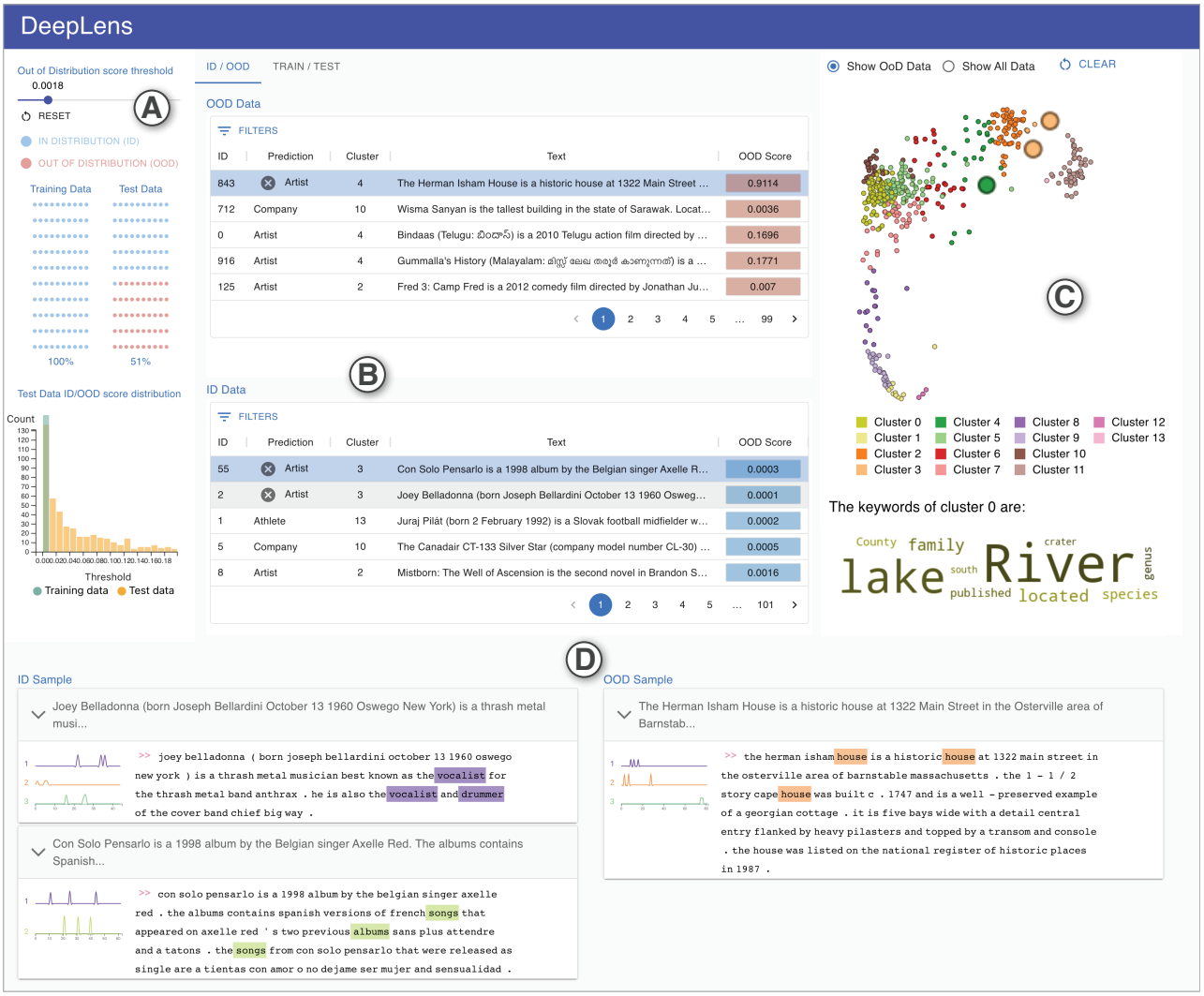


图2：DeepLens，一种用于检测和识别文本数据中分布外样本的交互式系统。(A) 分布视图 **允许用户调整阈值，并动态检查测试数据中的分布外问题。(B)** 实例视图 **在两个独立的交互式数据网格中显示分布内（ID）和分布外（OOD）数据。(C)**聚类视图 **显示聚类结果及每个簇的关键词，以探索潜在的分布外类型。(D)** 高亮视图 **在选定数据实例上显示高亮的显著词，以减轻用户的阅读努力程度。**

一个分布内实例，并查看两者的共性与差异（图 2 B◯）。此外，由于某些实例较长，DeepLens采用神经元活动分析 [23] 来识别并突出显示实例中的显著词，使用户能够快速掌握实例所涉及的主题（图 2 D◯）。通过直接比较分布外数据与分布内数据中高亮词语的差异，用户无需通读整个文本文件，便能轻松发现这两个实例之间潜在的主题差异。

为帮助用户识别不同类型的分布外数据（**N3**），DeepLens 会对检测到的分布外实例进行聚类，并将其呈现在散点图中（图2 C◯）。簇中的常见词被可视化为词云，以帮助用户理解其潜在的主题（图 2 C◯）。该聚类视图结合词云，可帮助用户从全局角度了解数据集中何时何地发生数据偏移（**N5**）。为支持（**N5**）中提到的“下钻”分析，DeepLens 允许用户通过点击某个簇中的节点或聚类图例，深入查看特定簇。实例视图将相应地被筛选。为支持（**N5**）中提到的“上卷”分析，DeepLens 会在聚类视图中突出显示用户选定的实例，使用户能够轻松查看所选实例在全局视图中的位置。最后，为支持**N4**，DeepLens 允许用户在实例视图 中按预测标签筛选实例，然后选择具有相同预测标签的分布外和分布内实例进行并排比较。显著词

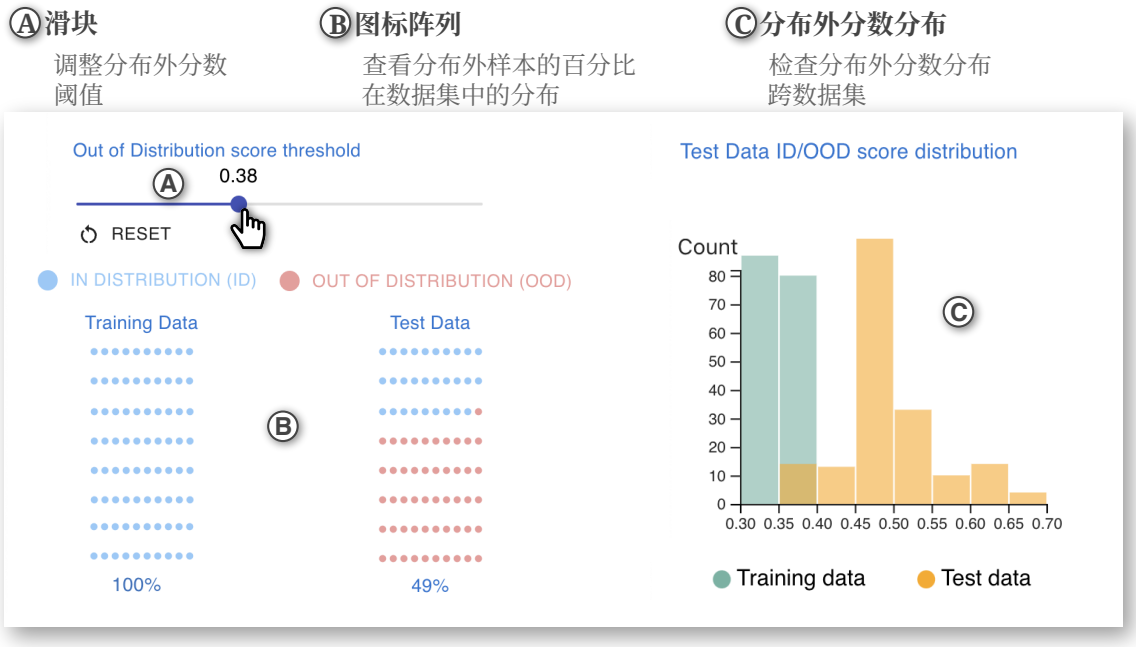


图 3：U 用户可以调整 分布外分数的阈值并检查 O OD 问题，在分布视图中 w.

高亮功能还可帮助用户快速了解分布外和分布内实例之间的共性与差异。

4 设计与实现

在本节中，我们遵循 DeepLens（图 1）的 3 步使用流程，介绍其设计与实现：（1）检测分布外问题，（2）探索分布外类型，以及（3）理解分布外数据。

4.1 交互式分布外文本检测

分布外检测方法。给定一个数据实例 x，分布外检测方法首先计算分布外分数 $s(x)$。如果 $s(x) > \epsilon$（预定义的阈值），则将 x视为分布外样本。DeepLens 采用基于校准的方法，即最大软最大概率 (maximum softmax probability) [21], 来计算分布外分数。较高的最大软最大概率表示模型对预测非常有信心，因此，较低的最大软最大概率意味着给定的数据实例 x 更有可能是分布外样本。给定一个概率分类器 C,

$$s(x) = 1 - \max_k C(y = k|x) \quad (1)$$

其中，$k \in 1$、.....、N表示类别标签 k，而 y表示 C的预测结果。请注意，即使自然语言处理模型并非用于分类，其中通常也存在概率分类器。例如，将隐藏状态投影到词汇表时，就存在概率分类器。

分布视图。 DeepLens 允许用户通过调整滑块，动态调整分布外检测的阈值（图 3 A◯）。更新阈值后，图标阵列 （图 3 B◯）也会相应更新。我们选择使用图标阵列，因为它能提供离散事件表示，并且已被证明比饼图和条形图等其他可视化方式更有助于准确解读数字，且所需的数感要求更低 [13, 28]。用户可以一眼查看图标阵列，快速了解测试数据中可能存在多少个分布外样本。为了验证是否设置了最优阈值，用户可以查看训练数据和测试数据中的 OOD 分数分布（图 3 C◯）。理想情况下，最优阈值应尽可能准确地区分分布内数据和分布外数据。DeepLens 提供了一个预先计算的阈值，帮助非专家高效地确定阈值。

实例视图。此视图包含两个可独立滚动的数据网格，分别用于显示分布内数据和分布外数据（图 4）。当用户更改分布外分数的阈值时，实例视图也会相应更新。每个数据网格的行表示单个数据实例，列包括每个数据实例的：索引、模型预测结果、聚类结果、原始文本和分布外分数 。默认情况下，包含分布外数据 和分布内数据 的数据网格分别按照分布外分数降序和升序排列。用户还可以筛选、搜索或排序每个数据网格，以探索数据实例。

4.2 分布外文本分类与探索

文本聚类。 为了帮助用户高效地探索检测到的分布外样本的主题，DeepLens 使用文本聚类算法对不同类型的文本进行分类。算法 1 描述了该聚类算法。给定一个自然语言处理模型 M，该算法首先提取新测试数据中每段文本的隐藏特征（第 1–5 行）。由于提取的特征 F 通常是稀疏且高维的向量，DeepLens 会应用 主成分分析 (主成分分析) [27] 将其维度降至 p （第 6 行）。然后，DeepLens 使用 K 均值聚类算法 [36] 对处理后的隐藏特征 F 进行聚类p （第 7–9 行）。为了确定最佳簇数n_{opt}，DeepLens 使用轮廓系数 方法 [53] （第 10 行）。在实现过程中，



图 4：与实例视图的交互。

算法 1： 一种语义文本聚类算法。

输入： 一个自然语言处理模型 M、新的测试数据 X、最大簇数 N_{max}、主成分分析维度 p

输出： 簇标签 Y

```
F ← {};
对于 x 在 X 中执行
    f ← 提取_特征(M, x);
    F .append(f);
结束
Fp ← 主成分分析(F p);
对于 n ← 1, . . ., N_max do
    Y_n = K 均值(Fp,n);
结束
n_opt ← 最大值_n 轮廓系数(Y_n, F_p);
返回 y_{n_opt};
```

我们将最大聚类数设为 $N_{max} = 200$，将主成分分析维度设为 $p = 128$。这些数值是根据经验确定的。

关键词摘要。确定每个簇后，DeepLens 会从单个簇中提取一些关键词并加以总结，帮助用户识别其潜在主题。为此，DeepLens 首先从每条数据中筛除"停用词" [6]，然后使用 CountVectorizer 算法 [48] 提取关键词。对于每个簇，DeepLens 会以词云形式展示出现频率最高的 10 个关键词。通过查看词云，用户可以快速了解所选簇可能包含哪些文本模式。

聚类视图。 DeepLens 将聚类结果和汇总的关键词整合到聚类视图中（图 5）。聚类视图由散点图和词云组成。散点图中的每个节点都代表一个独立的数据实例。每个节点的位置由主成分分析后各文本隐藏特征的前三个分量决定。每个节点的颜色代表其簇索引。当用户将鼠标悬停在某个节点上时，会弹出工具提示，显示相应数据实例的预测标签和分布外分数（图 5 B◯）。当用户点击某个节点时，相应的数据实例也会在实例视图中被选中。这些功能使用户能够结合具体的数据实例和文本来理解聚类结果。用户还可以点击图例，聚焦于某个簇（图 5 A◯）。选中某个簇后，DeepLens 会更新词云（图 5 C◯），并在实例视图中筛选掉所选簇之外的数据。

4.3 分布外文本解释

显著词选择。前文介绍了用户如何高效检查分布外问题并识别潜在的分布外类型。通过查看聚类视图，用户可能已经根据若干数据实例对潜在的分布外类型及其主题形成了假设。DeepLens 还支持用户深入分析特定实例。为此，DeepLens 使用神经元激活值分析 [23] 来选出文本中的显著词。通过查看小组高亮词语，用户无需仔细阅读冗长段落。算法 2 描述了显著词选择算法。总体而言，DeepLens 借助 ecco [1] 提取并分解神经元激活信息。对于给定文本 x，DeepLens 首先将其输入大型预训练语言模型 M，提取神经元激活值 A（第 2 行）。接着，DeepLens 使用非负矩阵分解，将提取的激活值分解为 n 个成分（第 3 行）。这样，DeepLens 就能将文本中的相似词语归为n 组。为了进一步降低用户的心理负担，我们筛除包含停用词或特殊标记（例如标点符号）的词组（第 4–14 行）。最后，对于每组词，DeepLens 只高亮显示激活值最高的 10 个词。这有助于仅保留组内最重要的词。在我们的实现中，我们使用了 HuggingFace [2] 发布的预训练 BERT [12]，且未进行任何微调。因子 n 数量设为 10，并且只高亮显示排名前 10 的显著词。

高亮视图。 在此视图中，DeepLens 利用 ecco 的可视化功能 [1] 来呈现高亮显示的显著词（图 6）。每个文本框左侧的迷你图展示了文本中相似显著词的位置。每个迷你折线图代表一组

[2]https://huggingface.co/bert-base-uncased

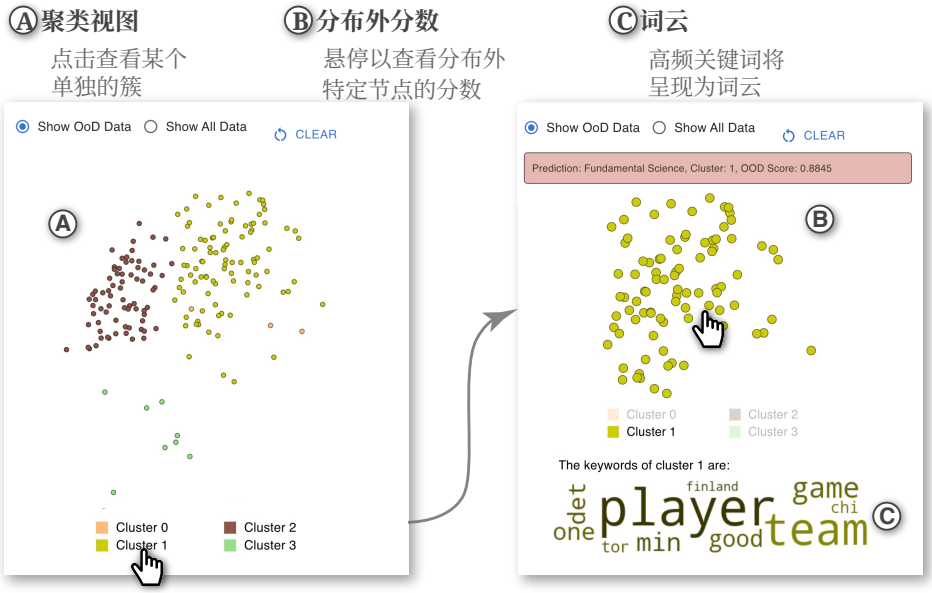


图 5：用户可以在聚类视图中检查并交互操作文本聚类结果。

算法 2： 显著词选择。

输入： 文本输入 x，预训练语言模型 M，因子数量 n ，文本中的词数 l
输出： 高亮关键词 S

$S \leftarrow \emptyset$;
$\mathcal{A} \leftarrow$ 获取_激活值(M, x);
$\tilde{\mathcal{S}} \leftarrow$ 非负矩阵分解(A n);
对于 $i \leftarrow 1, \ldots, n$;
  do
    对于 $j \leftarrow 1, \ldots, l$;
      do
        如果 x_j 是停用词或特殊标记 **那么**
          $S_i \leftarrow \emptyset$;
        else
          $\mathcal{S} \leftarrow S \cup \tilde{\mathcal{S}}_i$;
        end
    end
结束
返回 $\mathcal{S}$;

相似的显著词。迷你折线图的颜色与相应显著词组的颜色相同。横轴表示文本中词语的索引，纵轴表示词语的显著性分数。用户可以将鼠标悬停在不同的折线上，以查看不同组的显著词。对于每组显著词，颜色越深，表示激活值越高。此前的研究表明，通过对比方式帮助用户理解概念十分重要 [41]，DeepLens 支持用户同时固定多个分布内或分布外实例，以便对分布外主题进行情境化展示。当用户点击聚类视图 中的一个节点或在实例视图中的实例会固定显示在实例视图顶部。同时，每个选定实例的显著词会显示在高亮视图中。

4.4 实现

我们将 DeepLens 实现并部署为 Web应用。DeepLens 的界面（图 2）使用 Material UI [3]实现。我们使用 D3.js [4] 可视化散点图。所有机器学习模型均使用 PyTorch 和 Scikit-learn 实现，并在一块 NVIDIA A6000 GPU 上训练。为便于在用户研究期间访问，我们将 DeepLens 部署在 AWS EC2 上。

5 使用场景

假设 Alice 是一名模型开发者，她训练了一个机器学习模型，用于将文本文件分为两个不同的主题：信息技术（信息技术）和基础科学。她的模型在训练数据上的准确率达到 94%。然而，Alice 在线部署模型后，模型表现显著下降。Alice 怀疑，这种性能下降是由新数据中的分布外（OOD）问题造成的。Alice 在新数据上运行分布外检测程序，发现 49% 的新数据（400 个数据实例）被检测为分布外数据。Alice 想了解这些数据的特征，以便有针对性地应用数据增强技术来改进训练数据。然而，她发现手动检查这 400 个分布外样本并了解它们为何被归类为分布外样本非常耗时。此外，由于每份文本文件都很长，Alice 很难快速浏览并迅速掌握文档大意。

Alice 决定试用 DeepLens。她首先查看分布视图，其中两组图标阵列（图 3 B◯）展示了比例

[3]https://mui.com [4] https://d3js.org



Figure 6：高亮视图可帮助用户快速理解特定的分布内/分布外数据实例

分布内和分布外样本分别在训练数据和测试数据中的占比。她发现，与训练数据相比，测试数据中有很大一部分是分布外样本。她还查看了*OOD* 分数分布（图 3 C◯），其中训练和测试数据的 OOD 分数分布有所不同。她确认测试数据中存在分布外样本，这可能导致了她的模型性能下降。

Alice 想知道 DeepLens 检测到哪些类型的实例会被判定为分布外样本。因此，她转到实例视图 （图 2 B◯），其中分布外数据和分布内数据分别显示在两个独立的数据网格中。Alice 发现，分布外分数较高的实例非常可疑。由于逐一浏览每个数据实例会耗费太多时间，Alice 决定先切换到聚类视图 （图 2 C◯）查看概况。此视图包含一张散点图，其中每个数据实例都显示为一个节点，相似的数据实例会聚为一簇，并以相同颜色标示。默认情况下，聚类视图 仅显示分布外数据。Alice 发现，各簇中分布外节点的数量明显不同。接着，她将鼠标悬停在图例上，查看每个簇中分布外节点的确切数量。簇 1 和簇 2 中有许多分布外节点，而簇 0 和簇 3 中则少得多（图 5 A◯）。Alice 认为簇 1 和簇 2 可能包含两种新的分布外类型。

现在，Alice 想深入了解簇 1。点击相应图例后，聚类视图随即更新，只显示簇 1 的分布外节点（图 5 B◯）。实例视图 也会筛选掉不属于簇 1 的数据实例。词云 显示，簇 1 中有几个高频关键词，例如“球员”“比赛”“球队”（图 5 C◯）。Alice 意识到，这个簇的主题可能是体育。这明显表明分布外数据中出现了一个新类别，而该类别最初并未包含在训练数据中。

Alice 点击簇 1 中的一个节点。随后，选中的实例被固定在实例视图顶部。虽然选中的句子很长，需要花时间阅读，Alice 决定查看高亮视图 中的高亮词语（图 6 B◯）。这些显著词有助于她聚焦于关键信息，忽略句子中的无关词语。Alice 注意到有几个词语被高亮显示，例如“防守型”和“强打”。这些关键词进一步证实了 Alice 的判断：这是一个体育新类别。

为了进一步验证自己的判断，Alice 继续检查这个类别是否存在于分布内数据中。她点击实例视图中 ID 表的第一句话，该句被预测为基础科学 （图 6 A◯）。这句话中突出显示的显著词是：“医学”“结石”“缓解”和“治疗”（图 6 C◯）。这些词通常都出现在基础科学 文章中。接着，Alice 点击查看当前分布外实例附近的其他几个分布外实例，并确认体育 是分布外数据中的一个新主题。

6 用户研究

为评估 DeepLens 的有效性和易用性，我们对 24 名机器学习专业水平各异的程序员开展了一项被试内用户研究。为更好地了解 DeepLens 所提功能的价值，我们将 DeepLens 与一个禁用了聚类视图 和高亮视图的变体进行比较，并将该变体作为基线。

6.1 参与者

我们通过阿尔伯塔大学电气与计算机工程系和计算机科学系的邮件列表招募了 24 名参与者 [5]。所有参与者都具备机器学习的基础知识。10 名参与者是硕士生，10 名是博士生，3 名是专业开发人员，1 名是数据科学家。我们请参与者自行报告他们的机器学习专业水平。12 名参与者有 2 至 5 年的

[5]这项涉及人类参与者的研究已获该大学研究伦理办公室批准。

表 1：为用户研究设计的任务。

#	Task	描述	分布内数据	分布外数据	班次类型
1	主题分类	预测段落的主题 来自维基百科的文本。	DBPedia 顶部-4 [64]†	DBPedia 其余部分 [64]†	语义偏移
2	情感分析	预测一段文本的情感 评论文本。	IMDB [39]	Yelp [64]	背景偏移
3	情绪识别	识别情感 从给定文本中。	情感 负面-2 [54]‡	情感 其余部分 [54]‡	语义偏移
4	假新闻检测	检测一篇新闻文章是否 真实或虚假。	PolitiFact [55]	新冠肺炎 假新闻 [11] GossipCop [55]	背景偏移

† DBPedia 数据集有 14 个类别。我们将 DBPedia 前四类定义为按类别 ID 排序后包含前 4 个类别（公司、教育机构、艺术家和运动员）的子集，将 DBPedia 其余类别定义为包含其他 10 个类别的子集。‡ Emotion 数据集有 6 个类别。我们将 Emotion 负面两类定义为包含 2 个负面情感类别（悲伤和恐惧）的子集，将 Emotion 其余部分定义为包含其他 4 个类别的子集。

经验，1 名有 5 年以上，11 名只有 1 年经验。关于自然语言处理经验，5 名参与者有 2 至 5 年经验，19 名只有 1 年经验。20 名参与者表示，他们以前听说过分布外或分布偏移问题。所有研究环节均通过 Zoom 进行。DeepLens 和基线系统均以 Web 应用程序的形式部署，因此参与者可以使用自己的电脑参加研究环节。

6.2 任务

我们设计了四项任务，涵盖 NLP 领域中不同类型的分布偏移。表 1 展示了各项用户研究任务的详情。设计任务时，我们遵循以下要求：（1）任务应具有 NLP 领域的代表性；（2）任务应涵盖两种不同类型的分布偏移。为实现这些目标，我们从 NLP 分布外检测的先前研究以及自然语言处理模型的知名公开基准中收集了四项任务。对于每项任务，我们均采用一个 BERT 模型 [12] 作为骨干网络，并在分布内数据上微调模型。为确保公平比较，基线工具和 DeepLens 在每项任务中使用相同的预先计算分布外阈值。有关模型训练设置以及各项任务示例界面的更多详情，请参见附录 A。

6.3 实验流程

每次用户研究会话大约持续 60 分钟。在每次会话开始时，我们都会征得参与者同意进行录音。我们为参与者分配了两个用于识别分布外问题的任务：一个使用 DeepLens 完成，另一个使用基线工具完成。为减轻学习效应，我们在参与者之间平衡了任务分配顺序和工具分配顺序。总共有 6 名参与者体验了每个任务。在开始每项任务之前，参与者需要观看所分配工具的 5 分钟教程视频，随后进行 5 分钟的练习阶段，以熟悉该工具。然后，参与者有 20 分钟的时间使用所分配的工具，在给定的模型和数据集中检查并识别分布外问题。具体而言，参与者需要回答/报告：

(1) 它属于哪种数据分布偏移?
(2) 你发现了多少种不同类型的分布外数据?
（3）对于每种不同类型的分布外，请解释您认为其属于分布外的原因，并列出属于该类型的分布外实例的索引。

完成每项任务后，参与者填写任务后调查，就他们喜欢或不喜欢的方面提供反馈。参与者还需要回答五个 NASA 任务负荷指数（TLX）问题 [19]，作为任务后调查的一部分。完成两项任务后，参与者填写最终调查，直接比较所分配的两个工具。在用户研究会话结束时，每位参与者都会获得一张价值 25 美元的亚马逊礼品卡，以补偿他们付出的时间。

7 结果

本节报告并分析参与者使用 DeepLens 和基线工具时表现的差异。下文将参与者记为 P#。

7.1 用户表现

表 2 展示了参与者在识别分布外问题的四项任务中的表现。我们发现，使用 DeepLens 的全部 24 名参与者都正确识别出了指定模型中的数据分布偏移类型（即背景偏移与语义偏移），而使用基线方法的 24 名参与者中有 2 名未能正确识别。

为进一步评估参与者的表现，我们人工检查了参与者的回答，以核实其正确性。正确的分布外类型应当（1）包含能够代表一组数据实例的词语，且（2）与分布内数据显著不同。总体而言，与使用基线工具的参与者相比，使用 DeepLens 的参与者在全部四项不同任务中都找到了更多分布外类型。对于任务1和任务3（语义偏移），使用 DeepLens 的参与者平均分别找到了 7 种和 3.5 种分布外类型。相比之下，使用基线工具时，平均数量分别为 2.3 种和 1.2 种。Welch 的 t-检验表明，这两种情况下的表现差异均具有显著性（p-值 < 0.001）。对于任务2和任务4（背景偏移），大多数使用基线工具的参与者只能找到 1 种分布外类型（均值：1

表2：四项不同任务中的用户表现。

		任务1 语义偏移		任务2 背景偏移		任务3 语义偏移		任务4 背景偏移	
		基线	DeepLens	基线	DeepLens	基线	DeepLens	基线	DeepLens
正确的参与者人数识别出的班次类型		5	6	5	6	6	6	6	6
正确数量分布外类型每位参与者	Min	1/10	4/10	1/2	1/2	0/4	3/4	0/2	2/2
	Med	2/10	7.5/10	1/2	2/2	1/4	3.5/4	0.5/2	2/2
	Max	4/10	8/10	1/2	2/2	2/4	4/4	1/2	2/2
	Mean	2.3/10	7/10	1/2	1.7/2	1.2/4	3.5/4	0.5/2	2/2
Welch 的 t 检验		$\Delta = 4.7$, $t = 6.14$, $p < 0.001$		$\Delta = 0.7$, $t = 3.16$, $p = 0.010$		$\Delta = 2.3$, $t = 6.14$, $p < 0.001$		$\Delta = 1.5$, $t = 6.71$, $p = 0.001$	

种和 0.5 种）。相比之下，使用 DeepLens 的参与者在这两项任务中平均分别找到了 1.7 种和 2 种分布外类型。Welch 的 t-检验表明，表现差异具有显著性（p-值 = 0.010 和 p-值 = 0.001）。

我们分析了任务后调查的回答和录音，以了解使用 DeepLens 的参与者表现更好的原因。我们发现，DeepLens 用户的成功主要归功于聚类视图和高亮视图。首先，聚类视图显著加快了查找分布外类型的过程。24 名参与者中有 23 名大量使用聚类视图来探索分布外数据。相比之下，使用基线工具的参与者必须逐一检查分布外实例。P16 写道："*[使用基线工具时]逐一查看所有数据点很费时，尤其是数据点很多的时候。*"在任务后调查中，24 名参与者中也有 22 名认同聚类结果很有帮助。P17 表示："通过查看各个簇，我能比逐个查看数据点更快地发现趋势。"此外，每个簇的摘要关键词也被认为很有帮助。根据录音，24 名参与者中有 19 名从这些关键词开始探索。相比之下，使用基线工具的参与者通常会随机挑选一个实例作为探索起点。P20 表示："*[使用 DeepLens 时，]*我可以使用从簇中提取的关键词，并将其输入筛选器，以查找更多同类型的分布外实例。"P21 评论道："根据我的使用体验，*[DeepLens]*在快速归纳背景偏移关键词方面帮了我很多。"

此外，高亮视图帮助参与者避免了错误的分布外类型。在我们的用户研究中，使用 DeepLens 的参与者每人发现的错误分布外类型数量中位数为 0，而使用基线工具的参与者对应的数量为 1。错误分布外类型数量的均值差异为 0.52 对 1.64（Welch's t-检验：$p = 0.009$）。一个具体原因是，使用 DeepLens 时，参与者能够将分布外数据与分布内数据进行比较，以确认一种新的分布外类型。在任务后调查中，24 位参与者中有 17 位认为比较分布内数据和分布外数据很有帮助。此外，在比较分布内数据和分布外数据时，高亮关键词也帮助参与者避免误解长文本。在任务后调查中，24 位参与者中有 18 位同意，看到高亮关键词很有帮助。P9 表示："在验证我关于某个句子是否属于分布外数据的假设时，*DeepLens* 很有帮助，因为它会显示多个高亮关键词，并减少我花费的时间。"接下来

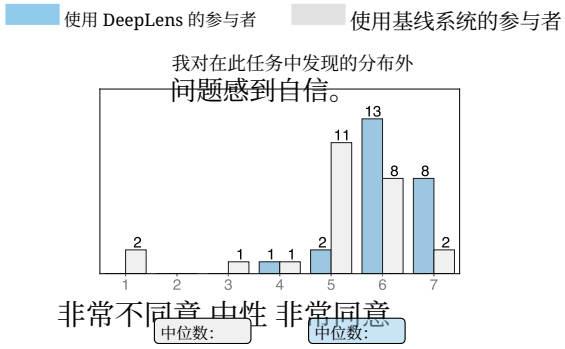


图7：用户对其任务表现的自我评价。

相比之下，P11 评论道："使用基线工具 *[*时，阅读整段文本并不容易*]*。"

我们还进一步分析了 OOD阈值调整功能对用户表现的影响。我们发现，只有两位用户（P6 和 P14）尝试调整过阈值，但最终都将其重置为预先计算的阈值。一个合理的解释是，默认阈值已经为用户调查分布外问题提供了良好的起点。因此，我们认为，分布视图中的 OOD 阈值可能对用户表现影响很小。由于两种条件下都提供了此阈值调整功能，DeepLens 更好的用户表现直接得益于聚类视图和高亮视图。

7.2 用户信心与认知负荷

在任务后调查中，参与者通过两个不同的7点李克特量表问题，自我报告了在指定工具的帮助下发现分布外问题时的信心。图7展示了参与者使用 DeepLens 和基线工具时的评价。我们发现，使用 DeepLens 的参与者对自己发现的分布外问题更有信心，其中信心评分的中位数分别为6和5。均值差为1.12（6.17和5.04），且在统计上显著（Welch's t-检验：p-值 = 0.002）。这种信心的提升主要归功于 DeepLens 的聚类视图。P5 评论道："自动聚类功能效果很好，而且关键词摘要非常有用，能帮助我全面了解情况。"P16 表示："通过对数据进行聚类，*DeepLens* 让词云中的关键词在识别分布外问题时更具指示性。"

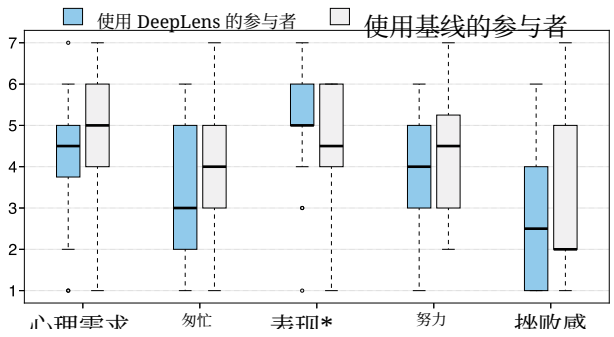


图8：NASA任务负荷指数评分。带星号 (*) 的条目表示均值差异具有统计显著性。

图8展示了参与者对 NASA TLX 问卷中五项认知因素的评分。尽管 DeepLens 功能更多、呈现的信息也更多，但我们发现，在心理需求、匆忙程度、努力程度和挫败感方面，使用 DeepLens 与基线工具之间没有显著差异（Welch's t-检验：$p = 0.22$, $p = 0.24$, $p = 0.26$, $p = 0.75$）。不过，与使用基线工具的参与者相比，使用 DeepLens 的参与者认为自己的表现更好（均值差：0.88，Welch's t-检验：$p = 0.049$）。这表明，与基线工具相比，DeepLens 在帮助用户检查机器学习模型并识别其中的分布外问题时，更有效且更有用。

7.3 个体功能的用户评分

在任务后调查中，参与者对 DeepLens 的关键功能进行了评分。在 24 名参与者中，23 人表示，在解决自己的机器学习模型中的分布外问题时愿意使用 DeepLens，另有 1 人持中立态度。中位数评分为 6 分（满分 7 点李克特量表；1 分表示"完全不想使用"，7 分表示"肯定会使用"）。如图9所示，参与者认为 DeepLens 的界面和交互功能直观且实用。参与者最喜欢聚类视图 。24 名参与者中有 22 人赞同"查看聚类结果很有帮助。"聚类视图的中位数评分为 6 分。P17 评论道："我喜欢它 *[DeepLens]* 具有聚类功能，例如，使用此功能可以非常迅速地找到'新冠'和'娱乐'分布外类别。" P20 评论道："*DeepLens* 的聚类和关键词可视化功能可以帮助我快速识别一种分布外类别。"此外，24 名参与者中有 18 人赞同"查看高亮关键词很有帮助。"高亮视图 的中位数评分为 6 分。P24 在任务后调查中评论道："*[DeepLens]* 让我少读一些内容，更容易专注于细节"24 名参与者中还有 17 人认为比较分布内和分布外数据很有帮助（中位数评分：6 分）。

7.4 用户偏好与反馈

在最终调查中，参与者自我报告了他们对 DeepLens 和基线工具的偏好（图 10）。24 名参与者中有 23 名表示 DeepLens 更有帮助（评分中位数：6），并且他们更愿意在实际工作中使用它（评分中位数：6）。我们对参与者对此问题的回答进行了编码，并确定了 2 个不同的主题。首先，17 名参与者提到，聚类视图让识别和分析分布外问题变得更容易。P23 评论道："关键词 *[在聚类视图中]* 面对大型数据集时很有用。" P13 表示："*[聚类视图]* 以可视化方式呈现数据，从而更容易识别分布外数据。"其次，5 名参与者将他们的成功归功于高亮视图。P9 表示："高亮关键词减少了我的耗时，也让我能轻松判断某个数据实例是否属于分布外。"

参与者还指出了 DeepLens 当前形式存在的一些局限性。2 位参与者表示，如果能改进 DeepLens 中的关键词摘要功能会更好。P21 说："有一些毫无意义的高频词，可能会造成干扰。"4 位参与者建议提升实例视图中筛选器的易用性，例如允许用户同时添加多个筛选条件。1 位参与者提到，语义词搜索功能（即匹配语义相似的词，而非完全相同的词）有助于他们进行探索。

8 讨论

8.1 设计启示

用户研究结果表明，与使用基线工具相比，借助 DeepLens，用户能够更有信心地发现更多类型的分布外数据。尽管解决分布外问题对于部署安全可靠的人工智能服务至关重要 [52]，，但大多数努力都投入到了提升分布外检测的算法准确率上。我们的工作表明，仅检测分布外样本并不足以提高模型开发者的工作效率，尤其是在数据集规模庞大且分布外类型多样的情况下。帮助开发者理解并探索大型文本语料库中的不同类型分布外数据，同样至关重要。开发者一旦深入了解其数据集中的分布外数据，就能进一步做出改善模型的战略决策，例如通过数据增强或筛选数据来重新训练模型。

在机器学习模型的持续交付过程中，新收集的数据量可能非常庞大。为了减少探索新数据中不同类型分布外实例所需的认知努力程度，必须为开发者总结出少量潜在的分布外类型。DeepLens 通过运用文本聚类算法来解决这一问题。此外，DeepLens 提供的交互式簇探索功能保留了用户对验证每种分布外数据类型的控制权。这符合其中一项人机交互指南——提供若干建议，而不是完全自动化这一流程 [2]。根据用户研究结果，我们发现，这种探索分布外数据的半自动化流程能够提升参与者的表现和信心。

虽然聚类结果无法直接告诉用户某种分布外类型究竟是什么，但每个簇的摘要关键词为探索提供了起点。在用户研究的最终研究调查中，23 位参与者中有 7 位表示更愿意在实际工作中使用 DeepLens，并明确提到摘要关键词对他们有所帮助。先前的研究 [9]表明，在设计用于图像数据分布外检测的交互式系统时，一个重要的要求是结合正常样本的上下文检查分布外样本。DeepLens 支持用户并排比较分布内文本和分布外文本。此外，与图像

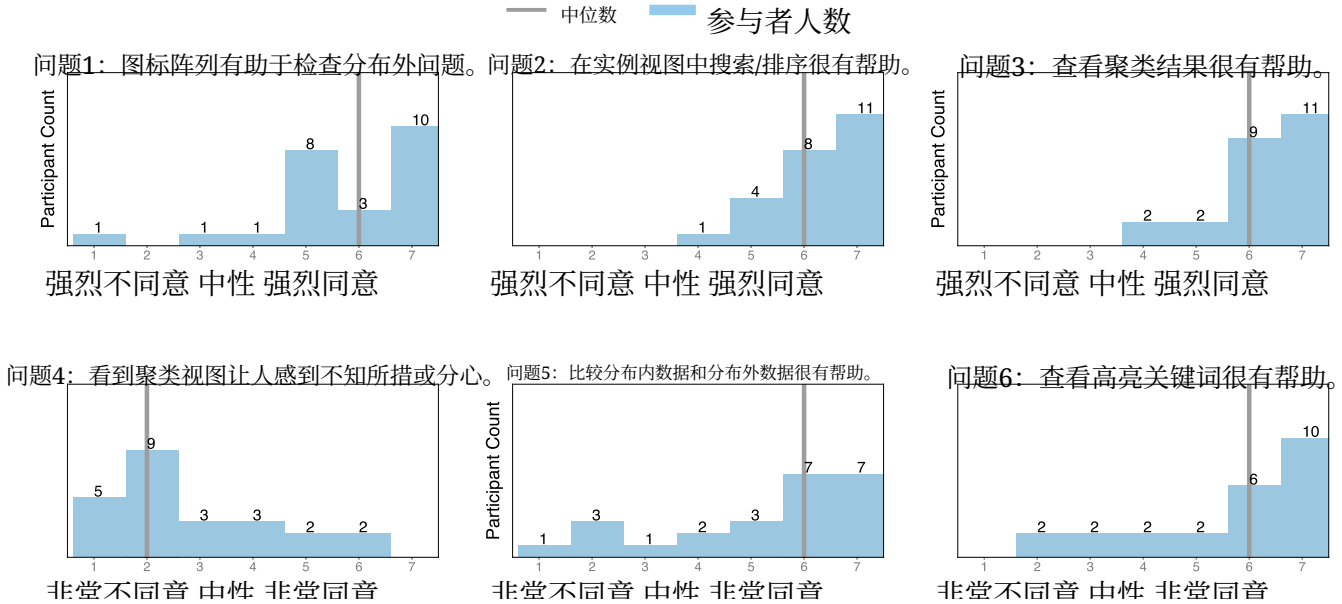


图 9：用户对 DeepLens 各项功能的评分。

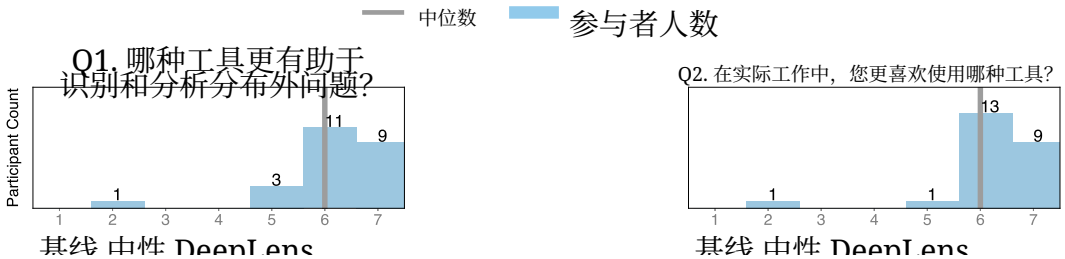


图 10：用户对分布外检测工具选择的偏好。

数据相比，文本文件不那么容易一眼浏览。因此，DeepLens 会突出显示每个文本文件中的显著词，帮助开发者快速掌握每份文档的要点。

8.2 目标用户群体和使用场景

DeepLens 专为了解机器学习但不熟悉分布外问题的用户设计。在我们的用户研究中，大多数参与者听说过分布外问题，但此前没有处理过这类问题。24 位参与者中有 4 位甚至表示，他们从未听说过分布外问题。用户研究结果表明，与使用基线工具相比，这些参与者使用 DeepLens 时表现更好（每位参与者发现的分布外类型均值：3.5 对 1.3，Welch's t-检验：$p < 0.0001$）。此外，他们对自己识别出的分布外问题也更有信心（信心评分中位数：6 对 5）。专家可能更倾向于通过阅读原始文本数据来识别分布外类型，但他们仍然认可 DeepLens，因为它能自动完成部分工作。例如，DeepLens 中的聚类功能可以自动将相似文本分类。P17 写道："通过查看这些簇，我能比逐个查看数据点更快地发现趋势。" P5 评论道："*DeepLens* 自动完成了一些手动工作，我发现自己的工作效率提高了。在相同的时间内，我可以完成更多工作。"

DeepLens 专为调试自然语言处理模型的分布外问题而设计。因此，我们的研究发现和设计启示无法推广到其他类型的机器学习问题，例如梯度消失。此外，DeepLens 也可以作为在线工具部署，以持续监控已部署模型中潜在的数据分布偏移。

8.3 局限性与未来工作

除了用户研究参与者指出的局限性和建议（第 7.4 节）之外，我们的用户研究设计和系统还存在其他一些局限性。

用户研究基线。在我们当前的用户研究中，通过禁用聚类视图和高亮视图，创建了 DeepLens 的一个变体作为基线方法。然而，这无法区分各项功能对用户性能提升的具体贡献。可以考虑通过禁用单个功能，创建更多 DeepLens 变体作为比较基线。

也可以考虑对该工具进行埋点，并在用户研究会话期间测量每项功能的使用效用率。

有限的自然语言处理任务。我们的用户研究无法确认 DeepLens 是否适用于所有类型的自然语言处理任务。为了全面评估 DeepLens 的实用性，可以考虑使用 DeepLens 在更多样化的自然语言处理任务中识别分布外问题，例如问答和自然语言推理。

分布外检测的准确率。目前，DeepLens 使用 最大软最大概率 [4] 作为分布外检测算法。尽管最大软最大概率已被证明在若干自然语言处理任务中有效 [4], ，但它未必总适用于其他类型的自然语言处理任务或模型。由于 DeepLens 的设计不局限于某一种分布外检测算法，未来的一个方向是将更多分布外检测算法集成到 DeepLens 中，并允许用户在不同算法之间切换。

可扩展性问题。根据我们的用户研究结果，DeepLens 可以处理 1,000 至 4,500 个数据点。然而，当数据规模扩大时（例如达到数百万个数据点），聚类视图中的节点可能会相互重叠。为了解决这一问题，可以采用更先进的可视化技术，例如气泡树图 [18] ，以分层方式呈现聚类结果。

替代算法与设计。我们的聚类视图可以通过采用更先进的降维和聚类算法得到进一步改进。在当前版本的 DeepLens 中，我们选择主成分分析进行降维，并选择 K-均值进行文本聚类，因为它们是经典且常用的选择。然而，更先进的降维方法，例如 t-SNE [24], Isomap [56] ，有望带来更好的降维结果。此外，我们的文本聚类还可以采用专门用于主题建模的方法加以改进，例如 ConceptScope [63] 和 TopicNets [16]。最后，我们的高亮视图也可以通过使用其他交互式工具来选择和可视化文本数据中的显著词，从而替代神经元激活分析，得到进一步改进，例如 exBERT [25]。

9 结论

本文介绍了一种新型交互式系统 DeepLens，旨在帮助机器学习开发者检测、探索并理解自然语言处理模型中潜在的分布外（OOD）问题。DeepLens 利用文本聚类算法，帮助用户高效地识别和探索大规模文本数据中潜在的分布外类型。此外，DeepLens 集成了一种基于神经元激活分析的算法，可突出显示单个数据实例中的显著词，帮助用户无需仔细阅读即可快速理解文本。我们将 DeepLens 实现为一款 Web应用，并针对四种不同的自然语言处理任务，与 24 名机器学习开发者开展了一项被试内用户研究。结果表明，与使用基线工具相比，在 DeepLens 的帮助下，开发者能够更好地理解机器学习模型中的分布外问题，并更有信心地识别出更多类型的分布外数据。最后，我们讨论了 DeepLens 带来的设计启示，并提出了若干值得探索的未来方向。

致谢

我们感谢用户研究中的所有匿名参与者以及匿名评审者提供的宝贵反馈。本研究部分得到了 Amii RAP 计划、加拿大 CIFAR AI 教席计划、加拿大自然科学与工程研究理事会（NSERC No.RGPIN-2021-02549、No.RGPAS-2021- 00034、No.DGECR-2021-00019）的支持，也得到了 JSPS 科研费助成金 No.JP20H04168 和 JST-Mirai 计划资助金 No.JPMJMI20B8 的支持。

参考文献

[1] J Alammar. 2021年. Ecco：一个用于解释 Transformer 语言模型的开源库。载于计算语言学协会第*59*届年会暨第*11*届自然语言处理国际联合会议论文集：系统演示。计算语言学协会。[2] Saleema Amershi、Dan Weld、Mihaela Vorvoreanu、Adam Fourney、Besmira Nushi、Penny Collisson、Jina Suh、Shamsi Iqbal、Paul N Bennett、Kori Inkpen 等。2019年。人机交互指南。载于*2019*年*CHI*计算机系统中的人为因素会议论文集。1–13。[3] Dario Amodei、Chris Olah、Jacob Steinhardt、Paul Christiano、John Schulman 和 Dan Mané。2016年。人工智能安全中的具体问题。*arXiv*预印本 *arXiv:1606.06565*（2016年）。[4] Udit Arora、William Huang 和 He He。2021年。分布外文本的类型及其检测方法。载于*2021*年自然语言处理实证方法会议论文集。10687–10701。[5] Koby Bibas、Meir Feder 和 Tal Hassner。2021年。用于分布外检测的单层预测归一化最大似然。神经信息处理系统进展34（2021年），1179–1191。[6] Steven Bird、Ewan Klein 和 Edward Loper。2009年。使用 *Python* 进行自然语言处理：借助自然语言工具包分析文本。“O’Reilly Media, Inc.”。[7] Julian Bitterwolf、Alexander Meinke 和 Matthias Hein。2020年。可认证的对抗鲁棒分布外数据检测。神经信息处理系统进展33（2020年），16085–16095。[8] Tom Bocklisch、Joey Faulkner、Nick Pawlowski 和 Alan Nichol。2017年。Rasa：开源语言理解与对话管理。*arXiv*预印本 *arXiv:1712.05181*（2017年）。[9] Changjian Chen、Jun Yuan、Yafeng Lu、Yang Liu、Hang Su、Songtao Yuan 和 Shixia Liu。2020年。Oodanalyzer：分布外样本的交互式分析。*IEEE*可视化与计算机图形学汇刊27, 7（2020年），3335–3349。[10] Sungik Choi 和 Sae-Young Chung。2020年。通过模糊进行新颖性检测。国际学习表征会议（*ICLR*）论文集（2020年）。[11] Sourya Dipta Das、Ayan Basak 和 Saikat Dutta。2021年。一种由启发式方法驱动的集成框架，用于新冠肺炎假新闻检测。载于紧急情况下区域语言网络敌意帖子治理国际研讨会。施普林格，164–176。[12] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018年。BERT：用于语言理解的深度双向 Transformer 预训练。*arXiv*预印本 *arXiv:1810.04805*（2018年）。[13] Mirta Galesic、Rocio Garcia-Retamero 和 Gerd Gigerenzer。2009年。使用图标阵列传达医疗风险：克服数感不足。健康心理学28, 2（2009年），210。[14] Ian Goodfellow、Jonathon Shlens 和 Christian Szegedy。2015年。解释并利用对抗样本。载于国际学习表征会议。http://arxiv.org/abs/ 1412 6572 。[15] Cheryl L Grady、Anthony R McIntosh、M Natasha Rajah 和 Fergus IM Craik。1998年。图片和词语情景编码的神经相关性。美国国家科学院院刊95, 5（1998年），2703–2708。[16] Brynjar Gretarsson、John O’donovan、Svetlin Bostandjiev、Tobias Höllerer、Arthur Asuncion、David Newman 和 Padhraic Smyth。2012年。Topicnets：利用主题建模对大型文本语料库进行可视化分析。*ACM*智能系统与技术汇刊（*TIST*）3, 2（2012年），1–26。[17] Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q Weinberger。2017年。现代神经网络的校准问题。载于国际机器学习会议。PMLR，1321–1330。[18] Jochen Görtler、Christoph Schulz、Daniel Weiskopf 和 Oliver Deussen。2018年。用于不确定性可视化的气泡树图。*IEEE*可视化与计算机图形学汇刊24, 1（2018年），719–728。https://doi.org/10.1109/TVCG. 2017.2743959[19] Sandra G Hart 和 Lowell E Staveland。1988年。NASA-TLX（任务负荷指数）的开发：实证与理论研究结果。载于心理学进展。第52卷。Elsevier，139–183。[20] Matthias Hein、Maksym Andriushchenko 和 Julian Bitterwolf。2019年。为何 ReLU 网络在远离训练数据的区域会产生高置信度预测

以及如何缓解这一问题。载于 IEEE/CVF计算机视觉与模式识别会议论文集。41–50。[21] Dan Hendrycks 和 Kevin Gimpel。2017年。神经网络中误分类样本和分布外样本检测的基线方法。载于第五届国际学习表征会议（ICLR 2017），法国土伦，2017年4月24日至26日，会议论文集。OpenReview.net。https://openreview.net/forum? id=Hkg4TI9xl [22] Dan Hendrycks、Norman Mu、Ekin D. Cubuk、Barret Zoph、Justin Gilmer 和 Balaji Lakshminarayanan。2020年。AugMix：一种提升鲁棒性和不确定性的简易数据处理方法。国际学习表征会议（ICLR）论文集（2020年）。[23] J Hewitt 和 P Liang。2019年。使用控制任务设计和解读探针。2019年会议论文集（2019年）。[24] Geoffrey E Hinton 和 Sam Roweis。2002年。随机邻域嵌入。神经信息处理系统进展15（2002年）。[25] Benjamin Hoover、Hendrik Strobelt 和 Sebastian Gehrmann。2019年。exbert：用于探索 Transformer 模型中学习到的表征的可视化分析工具。arXiv预印本 arXiv:1910.05276（2019年）。[26] Yen-Chang Hsu、Yilin Shen、Hongxia Jin 和 Zsolt Kira。2020年。广义ODIN：无需使用分布外数据进行学习即可检测分布外图像。载于 IEEE/CVF计算机视觉与模式识别会议论文集。10951–10960。[27] Ian T Jolliffe。2002年。特殊类型数据的主成分分析。施普林格。[28] Matthew Kay、Tara Kola、Jessica R Hullman 和 Sean A Munson。2016年。我的公交车什么时候（差不多）到？日常移动预测系统中以用户为中心的不确定性可视化。载于2016年CHI计算机系统中的人为因素会议论文集。5092–5103。[29] Jinhan Kim、Robert Feldt 和 Shin Yoo。2019年。利用惊讶充分性指导深度学习系统测试。载于2019年IEEE/ACM第41届软件工程国际会议（ICSE）。IEEE，1039–1049。[30] David Lazer、Ryan Kennedy、Gary King 和 Alessandro Vespignani。2014年。Google流感的寓言：大数据分析中的陷阱。《科学》343, 6176（2014年），1203–1205。[31] Kimin Lee、Kibok Lee、Honglak Lee 和 Jinwoo Shin。2018年。用于检测分布外样本和对抗攻击的简易统一框架。神经信息处理系统进展31（2018年）。[32] Rongjian Li、Wenlu Zhang、Heung-Il Suk、Li Wang、Jiang Li、Dinggang Shen 和 Shuiwang Ji。2014年。基于深度学习的影像数据补全，以改善脑疾病诊断。载于国际医学图像计算与计算机辅助干预会议。施普林格，305–312。[33] Shiyu Liang、Yixuan Li 和 R. Srikant。2018年。提升神经网络分布外图像检测的可靠性。载于国际学习表征会议。https://openreview.net/forum?id=H1VGkIxRZ [34] Ziqian Lin、Sreya Dutta Roy 和 Yixuan Li。2021年。MOOD：多层次分布外检测。载于 IEEE/CVF计算机视觉与模式识别会议论文集。15313–15323。[35] Weitang Liu、Xiaoyun Wang、John Owens 和 Yixuan Li。2020年。基于能量的分布外检测。神经信息处理系统进展33（2020年），21464–21475。[36] Stuart Lloyd。1982年。PCM中的最小二乘量化。IEEE信息论汇刊28, 2（1982年），129–137。[37] Christos Louizos 和 Max Welling。2017年。用于变分贝叶斯神经网络的乘性归一化流。载于国际机器学习会议。PMLR，2218–2227。[38] Jie Lu、Anjin Liu、Fan Dong、Feng Gu、Joao Gama 和 Guangquan Zhang。2018年。概念漂移下的学习：综述。IEEE知识与数据工程汇刊31, 12（2018年），2346–2363。[39] Andrew Maas、Raymond E Daly、Peter T Pham、Dan Huang、Andrew Y Ng 和 Christopher Potts。2011年。用于情感分析的词向量学习。载于计算语言学协会第49届年会：人类语言技术论文集。142–150。[40] Alexander Meinke 和 Matthias Hein。2020年。迈向能够以可证明方式知晓自身所不知的神经网络。国际学习表征会议（ICLR）论文集（2020年）。[41] Tim Miller。2019年。人工智能中的解释：来自社会科学的见解。《人工智能》267（2019年），1–38。[42] Mehryar Mohri、Afshin Rostamizadeh 和 Ameet Talwalkar。2018年。机器学习基础。麻省理工学院出版社。[43] Jose G Moreno-Torres、Troy Raeder、Rocío Alaiz-Rodríguez、Nitesh V Chawla 和 Francisco Herrera。2012年。分类中数据集偏移的统一观点。《模式识别》45, 1（2012年），521–530。[44] Peyman Morteza 和 Yixuan Li。2022年。理解分布外检测的可证明保证。载于AAAI人工智能会议论文集，第8卷。[45] Anh Nguyen、Jason Yosinski 和 Jeff Clune。2015年。深度神经网络很容易受骗：对不可识别图像的高置信度预测。载于 IEEE计算机视觉与模式识别会议论文集。427–436。

[46] Matthew L Olson、Thuy-Vy Nguyen、Gaurav Dixit、Neale Ratzlaff、Weng-Keen Wong 和 Minsuk Kahng。2021年。图像数据中协变量偏移的对比识别。载于*2021 IEEE*可视化大会 *(VIS)*。IEEE，36–40。[47] João Palmeiro、Beatriz Malveiro、Rita Costa、David Polido、Ricardo Moreira 和 Pedro Bizarro。2022年。Data+ Shift：支持数据科学家对数据分布偏移进行可视化调查。*arXiv*预印本 *arXiv:2204.14025*（2022年）。[48] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cour-napeau、M. Brucher、M. Perrot 和 E. Duchesnay。2011年。Scikit-learn：Python中的机器学习。机器学习研究杂志12（2011年），2825–2830。[49] Zi Peng、Jinqiu Yang、Tse-Hsun Chen 和 Lei Ma。2020年。初探机器学习模型在复杂自动驾驶系统中的集成：Apollo案例研究。载于第*28*届*ACM*欧洲软件工程会议与软件工程基础研讨会联合会议论文集。1240–1250。[50] Alise J Ponsero 和 Bonnie L Hurwitz。2019年。机器学习在水生宏基因组中检测病毒的前景与陷阱。《微生物学前沿》10（2019年），806。[51] Stephan Rabanser、Stephan Günnemann 和 Zachary Lipton。2019年。大声告警：检测数据集偏移方法的实证研究。神经信息处理系统进展32（2019年）。[52] Jie Ren、Peter J Liu、Emily Fertig、Jasper Snoek、Ryan Poplin、Mark Depristo、Joshua Dillon 和 Balaji Lakshminarayanan。2019年。用于分布外检测的似然比。神经信息处理系统进展32（2019年）。[53] Peter J Rousseeuw。1987年。轮廓系数：用于解释和验证聚类分析的图形辅助工具。《计算与应用数学杂志》20（1987年），53–65。[54] Elvis Saravia、Hsien-Chi Toby Liu、Yen-Hao Huang、Junlin Wu 和 Yi-Shin Chen。2018年。Carer：用于情绪识别的情境化情感表征。载于*2018*年自然语言处理实证方法会议论文集。3687–3697。[55] Kai Shu、Deepak Mahudeswaran、Suhang Wang、Dongwon Lee 和 Huan Liu。2020年。FakeNewsNet：一个包含新闻内容、社会背景及时空信息的数据存储库，用于研究社交媒体上的假新闻。《大数据》8，3（2020年），171–188。[56] Joshua B Tenenbaum、Vin de Silva 和 John C Langford。2000年。用于非线性降维的全局几何框架。《科学》290，5500（2000年），2319–2323。[57] Sunil Thulasidasan、Gopinath Chennupati、Jeff A Bilmes、Tanmoy Bhattacharya 和 Sarah Michalak。2019年。关于Mixup训练：改进深度神经网络的校准和预测不确定性。神经信息处理系统进展32（2019年）。[58] Haoran Wang、Weitang Liu、Alex Bocchieri 和 Yixuan Li。2021年。多标签分类网络能否知道自己不知道什么？神经信息处理系统进展34（2021年），29074–29087。[59] Xumeng Wang、Wei Chen、Jiazhi Xia、Zexian Chen、Dongshi Xu、Xiangyang Wu、Mingliang Xu 和 Tobias Schreck。2020年。ConceptExplorer：多源时间序列数据中概念漂移的可视化分析。载于*2020*年*IEEE*可视分析科学与技术会议 *(VAST)*。IEEE，1–11。[60] Weikai Yang、Zhen Li、Mengchen Liu、Yafeng Lu、Kelei Cao、Ross Maciejewski 和 Shixia Liu。2020年。通过可视分析诊断概念漂移。载于*2020*年*IEEE*可视分析科学与技术会议 *(VAST)*。IEEE，12–23。[61] Anton Yeshchenko、Claudio Di Ciccio、Jan Mendling 和 Artem Polyvyanyy。2021年。用于业务流程序列数据分析的可视化漂移检测。*IEEE*可视化与计算机图形学汇刊（2021年）。[62] Sangdoo Yun、Dongyoon Han、Seong Joon Oh、Sanghyuk Chun、Junsuk Choe 和 Youngjoon Yoo。2019年。CutMix：利用可定位特征训练强分类器的正则化策略。载于*IEEE/CVF*国际计算机视觉会议论文集。6023–6032。[63] Xiaoyu Zhang、Senthil Chandrasegaran 和 Kwan-Liu Ma。2021年。ConceptScope：基于领域本体组织并可视化文档中的知识。载于*2021*年*CHI*计算机系统中的人为因素会议论文集。1–13。[64] Xiang Zhang、Junbo Zhao 和 Yann LeCun。2015年。用于文本分类的字符级卷积网络。神经信息处理系统进展28（2015年）。

A 用户研究任务

A.1 自然语言处理任务1：主题分类

DBPedia 数据集从维基百科项目创建的信息中提取结构化内容。在此任务中，我们使用 DBpedia-14，该数据集于2014年通过从维基百科中选取14个互不重叠的主题构建而成。我们使用前4个类别中的样本作为分布内数据，而

其余部分作为分布外数据。为模拟具有语义偏移的真实世界数据，我们从14个主题的测试集划分中抽取1000个实例。
分布内数据包含 300 个样本和 4 个主题：公司、教育机构、艺术家和运动员。
分布外数据包含 700 个样本和 10 个主题：公职人员、交通工具、建筑物、自然地点、村庄、动物、植物、专辑、电影和书面作品。

模型我们使用分布内数据的训练集划分，以 5×10^{-5} 的学习率和 16 的批次大小对 BERT 模型进行 1 个周期的微调。该模型在分布内数据的验证集划分上达到 98.5% 的准确率，在测试数据上达到 28% 的准确率。

A.2 自然语言处理任务2：情感分析

在此任务中，我们使用 IMDB 数据集作为分布内数据，并使用 Yelp Polarity 二元情感分类数据集作为分布外数据。IMDB 数据集包含电影评论。Yelp Polarity 数据集由针对不同商家的评论构成。IMDB 和 Yelp 数据集都包含两个标签：正面和负面，用于预测评论的情感。为创建具有背景偏移的在线数据，我们从 IMDB 和 Yelp 测试集划分中各采样 1000 个实例。
分布内数据 来自 IMDB 数据集的 495 条电影评论样本。

分布外数据 来自 Yelp 数据集的 505 条商业评论样本。

模型 我们在分布内数据的训练集划分上对 BERT 模型进行 1 轮微调，学习率为 5×10^{-5}，批次大小为 16。该模型在分布内数据的验证集划分上达到 93.5% 的准确率，在测试数据上达到 89% 的准确率。

A.3 自然语言处理任务3：情感识别

情感数据集包含六种基本情绪：悲伤、恐惧、喜悦、愤怒、惊讶和爱。该数据集来源于英语推特消息。数据集包含两列，分别对应情绪索引（0 到 5）和文本。我们将悲伤类和恐惧类的样本作为分布内数据，其余部分作为分布外数据。为了创建具有语义偏移的在线数据，我们从 6 个主题的测试集划分中抽取 1000 个实例。

分布内数据 分布内数据包含 644 个样本和 2 种情感：悲伤和恐惧。

分布外数据 分布外数据包含 356 个样本和 4 种情感：喜悦、愤怒、惊讶和爱。
模型 我们使用学习率 2×10^{-5} 和批次大小 32，在分布内数据的训练集划分上对 BERT 模型进行 4 轮微调。该模型在分布内数据的验证集划分上达到 99.2% 的准确率，在测试数据上达到 31% 的准确率。

A.4 自然语言处理任务4：假新闻检测

在此任务中，我们设计了一种背景偏移场景，涉及对不同类型新闻进行虚假性检测。FakeNewsNet 是一个从两个事实核查网站 GossipCop 和 PolitiFact 收集的数据集，其中包含新闻及其有效性标签，这些标签由专业记者和专家标注。PolitiFact 包含与美国政治相关的新闻，而 GossipCop 则由娱乐新闻和八卦新闻组成。此外，我们还在此任务中使用 COVID-19假新闻数据集，其中包含从 Facebook、Twitter 等社交媒体平台提取的 COVID-19相关新闻。我们将 PolitiFact 数据作为分布内数据，然后将 GossipCop 和 COVID-19假新闻数据集合并为分布外数据。为了创建具有背景偏移的在线数据，我们从 PolitiFact、GossipCop 和 COVID-19假新闻数据集的测试集划分中抽取 4500 个样本。
分布内数据 2000年篇与美国政治相关的新闻。**分布外数据** 2500 篇与八卦和新冠肺炎相关的新闻。

模型 我们使用学习率为 5×10^{-5}、批次大小为 16，在分布内数据的训练集划分上对 BERT 模型进行 3 个周期的微调。该模型在分布内数据的验证集划分上达到 89.5% 的准确率，在测试数据上达到 67% 的准确率。

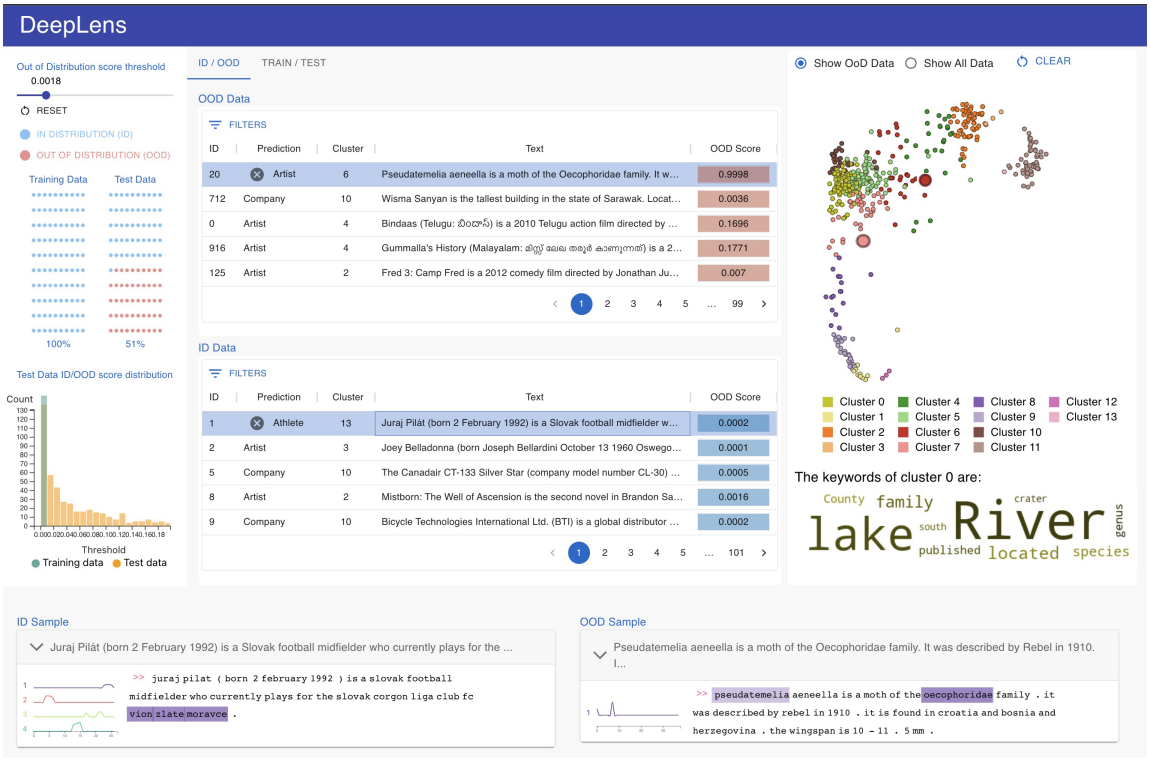
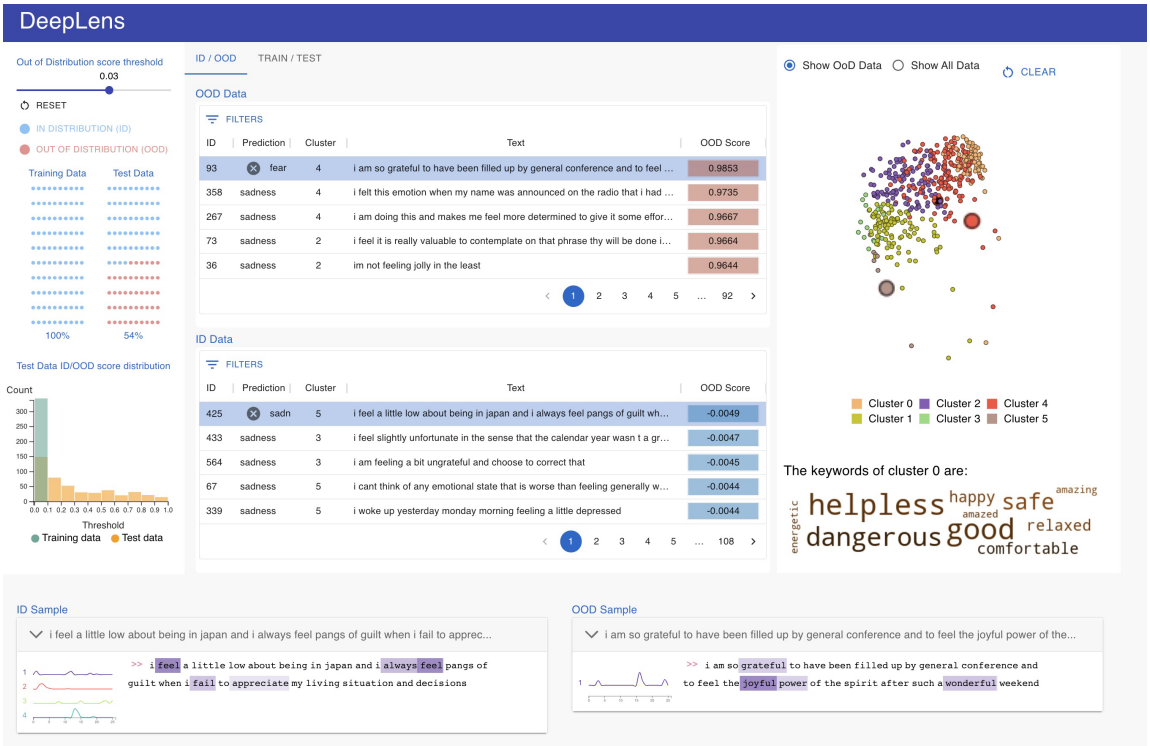


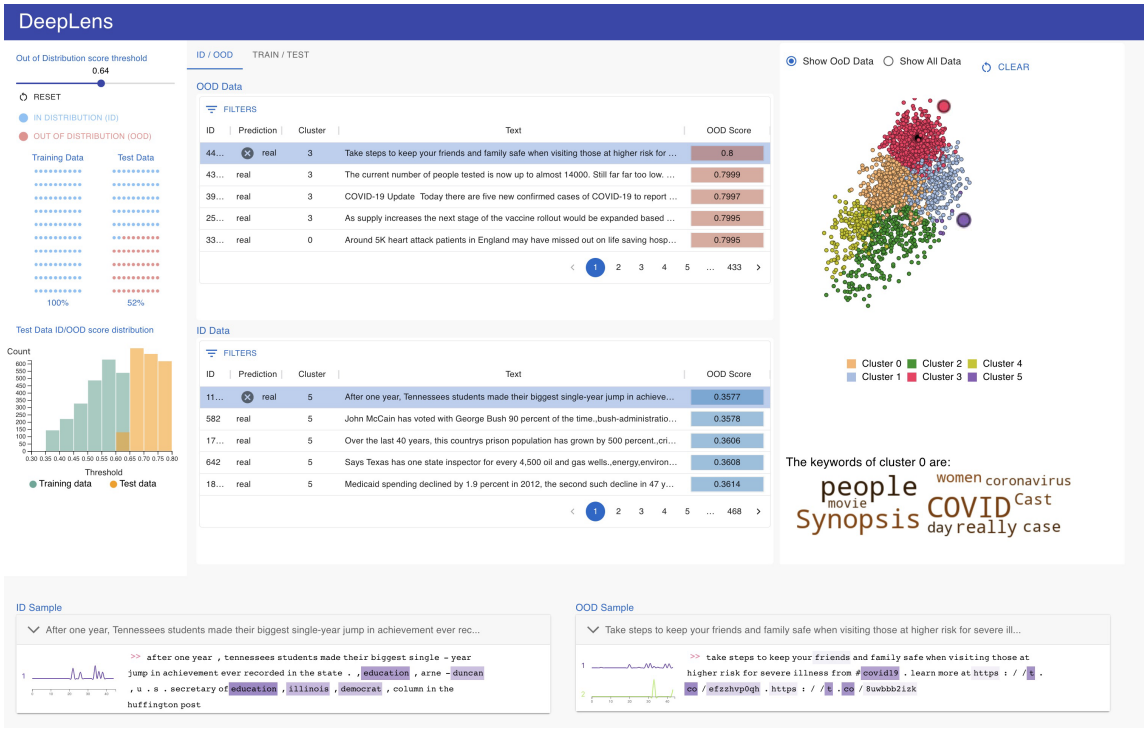
图 A1：DeepLens 任务1界面。



图 A2：DeepLens 任务2界面。



图A3：DeepLens 任务3界面。



图A4：DeepLens 任务4界面。